Data types

Measurement scales

- **Nominal:** labels only; mode and counts.
- **Ordinal:** ranked labels; median and percentiles.
- **Interval:** equal gaps, no true zero; differences valid.
- **Ratio:** equal gaps and true zero; ratios valid.

Variables

- **Discrete:** countable values.
- **Continuous:** values on an interval.
- **Missing:** record why; never silently replace with zero.

Descriptive statistics

Centre

- Mean: $\bar{x} = n^{-1} \sum_i x_i$.
- Median: middle ordered observation.
- Weighted mean: $\bar{x}_w = \sum_i w_i x_i / \sum_i w_i$.

Spread

- Range: $\max(x) - \min(x)$.
- Variance: $s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$.
- Standard deviation: $s = \sqrt{s^2}$.
- IQR: $Q_3 - Q_1$; robust to outliers.

Useful checks

- Compare units before comparing magnitudes.
- Plot the distribution before choosing a summary.
- Report centre together with a spread measure.

Probability

Rules

- Complement: $P(A^c) = 1 - P(A)$.
- Union: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- Conditional: $P(A \mid B) = P(A \cap B)/P(B)$.
- Independence: $P(A \cap B) = P(A)P(B)$.

Bayes

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

- **Prior:** belief before observing data.
- **Likelihood:** data probability under a model.
- **Posterior:** updated belief after data.

Random variables

Expectation

- $E[X] = \sum_x x P(X = x)$ for discrete X.
- $\mathrm{Var}(X) = E[X^2] - E[X]^2$.
- $E[aX + b] = aE[X] + b$.
- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.

Common distributions

Model	Mean; variance
Bernoulli(p)	p; $p(1-p)$
Binomial(n, p)	np; $np(1-p)$
Poisson(λ)	λ; λ
Normal(μ, σ^2)	μ; σ^2

Sampling

Standard error

- $SE(\bar{x}) = s/\sqrt{n}$.
- Larger samples reduce standard error, not necessarily bias.
- Central limit approximation: $\bar{x} \approx N(\mu, \sigma^2/n)$.

Intervals

- Approximate 95% mean interval: $\bar{x} \pm 1.96\,SE$.
- A confidence procedure has coverage over repeated samples.
- State estimator, confidence level, and assumptions.

Sampling bias

- **Coverage:** target population missing from frame.
- **Non-response:** respondents differ systematically.
- **Measurement:** instrument or wording changes answers.

Hypothesis tests

Workflow

1. State H_0, H_1, parameter, and direction.
2. Choose α and a test statistic.
3. Compute statistic and p-value.
4. Report effect size and uncertainty, not p-value alone.

Errors

- Type I: reject a true null; probability α.
- Type II: fail to reject a false null; probability β.
- Power: $1 - \beta$; increases with effect and sample size.
- A p-value is not $P(H_0 \mid \text{data})$.

Decision

- If $p \le \alpha$, reject H_0.
- If $p > \alpha$, do not reject H_0.
- Statistical significance is not practical significance.

Regression

Linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad E[\varepsilon_i \mid x_i] = 0.$$

- β_1: expected change in y for one unit of x.
- $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$.
- Residual: $e_i = y_i - \hat{y}_i$.

Assumptions

- Linearity, independent errors, constant variance.
- Normal errors support small-sample inference.
- Inspect residuals versus fitted values and predictors.
- Correlation does not establish causation.

Classification

- Precision $= TP/(TP + FP)$; recall $= TP/(TP + FN)$.
- Specificity $= TN/(TN + FP)$.
- F1 $= 2PR/(P + R)$.

Data workflow

Prepare

1. Define the unit of analysis and target.
2. Validate types, ranges, duplicates, and missingness.
3. Split train/test before fitting transformations.

Evaluate

- Use a baseline before a complex model.
- Choose metrics that reflect the decision cost.
- Cross-validation estimates out-of-sample performance.
- Keep a final test set untouched until the end.

Reproducibility

- Record data version, code version, seed, and environment.
- Separate exploratory analysis from confirmatory claims.
- Preserve transformations with the fitted model.

Exam checklist

Before answering

- Define symbols and units.
- Draw the relevant quantity or data-generating process.
- Write the assumptions before applying a formula.
- Check limiting cases and sign conventions.
- Round only at the final step.

Interpretation

- Give the direction, magnitude, and uncertainty.
- Distinguish association, prediction, and causation.
- Explain what the result does not show.

Workspace